Y Nut, a Phonetic-based Learning System for Spoken Languages

Omer Landry Nguena Timo, Tegawendé F. Bissyandé

Y Nut, a Phonetic-based Learning System for Spoken Languages

Omer L. Nguena Timo[1], Tegawendé F. Bissyandé[2]

[1] LaBRI, University of Bordeaux - CNRS, France
nguena@labri.fr
[2] SnT, University of Luxembourg, Luxembourg
tegawende.bissyande@uni.lu

Abstract. Communication between humans is of importance for our societies. It requires constant learning of new languages, e.g., by travellers whose extended stay in foreign locations facilitate learning. When a language possesses a written form, much of the meaning necessary for its learning is directly provided by the text. In spoken languages however, the meaning is only vehicled by the sounds. Nonetheless, learning spoken languages can take advantage of linguistic contents available in audio or video media which abound on the Internet and the social networks. We open a discussion and describe a system that enables to enrich a phonetic database so as to ease learning of basic expressions of spoken languages. Such a system could be useful for the survival of the plethora of spoken languages in Africa. The purpose of such a system is to provide within a reasonable period, automatic syntactic translation services.

1 Introduction

By the year 2100, 50 to 90% of the current 7,000 languages spoken across the world will be extinct [1]. Recent statistical summaries[1] have reported that most of the still alive languages are actually spoken in developing regions in Africa (30%) and Asia (30%) where the loss of *spoken languages*, aka *oral languages*, will actually significantly harm cultural diversity. Today, spoken languages in african developing areas are endangered because of the cultural/political/economic hegemony of the West [1]. Nonetheless, there is currently a momentum of (re)learning languages that estranged populations now value in various contexts. For example, a song in *Lingala*[2] broadcasted over the internet may spark the interest of an expatriate congolese, while old recordings of historical speeches by african elders may need to be valued.

Thanks to video-sharing sites and web-based music streaming services, an abundance of multi-media content now proliferaes on the internet. Because it originates from all parts of the world, this content features a considerable number of spoken languages. And because the internet can be accessed from all around

[1] http://www.ethnologue.com/statistics/area
[2] Bantu language spoken in Congo

the globe, we are presented with ample opportunity to leverage its multimedia content to learn languages in an adequate socio-technical environment.

It is commonly accepted that the difficulty for learning a language is correlated to the difficulty for the learner to recognize the sounds, i.e phones [3], made when speaking the language. Furthermore, in a given language, several phones can actually be perceived as equivalent although they are not identically pronounced (e.g., the 'k' sound is apirated in the word 'kit' and unaspirated in 'skill'). Such a set of phones is known as a *phoneme* and will be transcribed identically. As a result, transcribing phonemes of a given language is an important step towards providing a corpus for its learning.

State of the art In previous work, Nguyen Thi Minh has studied the phonetic transcription of Vietnamese [4]. This work was facilitated by the fact that Vietnamese is a written language and already possesses an alphabet. Purely spoken languages present distinct challenges for phonetic transcriptions. Our endeavour however involves the transcription of such languages which are the majority of endangered languages in developing regions. Work on phonetic transcriptions has mainly relied on two basic techniques :

- **Rule-based approaches:** such approaches require the implementation of transcription rules, and are thus adapted for languages that are already equipped with such rules. The simpler and precise the rules are, the faster and precise the transcription is. Ordean *et al.* have implemented *Grapheme-to-Phoneme* trancription rules that were integrated in the text processing component of a text-to-speech system for Romanian [5].
- **Statistical approaches based on learning:** These approaches are used with an initial lexicon that is progressively enriched by applying statistical techniques to infer new transcriptions. This method is particularly adapted for the extension to new languages. Besling has also used a statistical approach to derive phonetic transcriptions independantly of the languages, building upon background lexica [2].

This paper. We propose in this paper the Ynut system for providing a socio-technical environment for the transcription of spoken languages. The contributions of this paper are:

1. we introduced a new research avenue for african researchers in the learning of spoken languages using multimedia content.
2. we propose a collaborative approach to learning, which is likely to benefit from the cultural model of Africa [6].
3. we discuss the design of the Ynut system for enriching the huge amount of multimedia content on the internet with phonetic transcriptions that will help learning.

The remainder of this paper is structured as follows. Section 2 introduces domain-specific terms that are necessary to understand our approach. Section 3 then details the Ynut system, describing each module and what service it delivers. We conclude in Section 4.

2 Preliminaries

This section introduces the preliminaries to understand the scope of our research project, and how realistic our endeavour is. We first detail different concepts that are necessary to grasp the challenges of phonetic transcriptions. Then, we discuss the abundance of multimedia content and the opportunities that they provide.

Phonetic systems and units. `Phonemes` are the basic linguistic units of languages' phonology. Substituting a phoneme for another one modifies the listening and the meaning of the speech. In written languages, each phoneme can be represented with a combination of symbols of an `alphabet`. Each symbol of an alphabet is also called a `grapheme`. One grapheme also represents one phoneme. Written languages are equipped with `dictionaries` that enumerate the `root words` of the languages. Root words are represented with sequences of graphemes. Written languages also provide rules for creating and composing new valid words and phrases.

Multimedia broadcasting services. With the latest advances on data storage systems and the improvement of Internet connection across the world, multimedia broadcasting services such as Youtube, Dailymotion and Netflix for video or Deezer for audio, the traffic on the Internet mostly deals with multimedia content. Social networking paradigm has anabled user experience of videos to be more inclusive by allowing a single video clip to be passed over and shared by thousands in a matter of days. Such epidemic spread of multimedia content, their lifetime on the web, and their random origins from anywhere on earth, make them a very good medium to assess the flourishing of various spoken languages. Unfortunately, listening and watching videos is even more useful for a non-native of the spoken language when there are accompanying subtitles. However writing subtitles for a multimedia content is a daunting, solitary, and sometimes unrewarding task. For spoken languages, the challenges are aggravated by the limited number of subtitles that must be transcribed in a foreign alphabet.

3 The Ynut System

Ynut is a pilot project for a more effective learning of spoken languages leveraging multimedia content that is available on the Internet and that potential learners already stream everyday. The Ynut system thus aims at enabling users to enrich audio and video media with transcription items refered to as *yphemes*. Yphemes are representations of morphemes in the Ynut systems. A ypheme is based on written languages and each instance of ypheme obeys to the phonetic system of the associated written language. It allows us to represent speeches of a spoken language. Intuitively, since spoken languages are not equipped with dictionaries, people that have rudiment (approximative or exact writing) knowledge about written languages can represent the speeches in spoken languages with morphemes of those written languages. The Ynut system is compromised

of five modules, each delivering a specific service that will contribute to improve collaborative learning of spoken languages. As a collaborative system, Ynut relies on users to contribute with the yphemes for spoken languages and the associated written language that they understand.

3.1 The yphemes Recording Module

The first module of the Ynut system is a recording Module for construction of a knowledge database. The yphemes Recording Module is in charge of recording yphemes provided by a contributor. Before he starts recording new yphemes for a given spoken language, the contributor must associate an input written language, (a part of) video/audio media, the translations of the yphemes into another written language or optionally into the international phonetic system. Thus the created corpus will contain information for linking sounds from a spoken language and their transcriptions for reading into a written language.

3.2 The Phonetic Transcription Module

The second module is internal to Ynut and does not require interaction with users. It is a phonetic transcription module which automatically transcripts recorded yphemes into the phonetic system of the associated written language or the international phonetic system. Example 1 illustrates an example of translation between Ynut's yphemes and the Internation Phonetic Alphabet (IPA).

Example 1. Here is a translation of two yphemes into the English IPA[3].

ypheme	**IPA**
y nut	ˈwaɪˈnət
why not	ˈwaɪˈnɑt

We observe from the above example that the two IPA phonemes are very similar although the yphemes look more different, thus highlighting how the Ynut systems addresses the challenge having contributors recording distinct yphemes for the same heard sound.

3.3 The Search Module

The Search module uses information retrieval techniques to enable the search for for yphemes or words according to similarity criteria. In Ynut, these similarity criteria may include syntactic similarity between the yphemes or phonetic similarity.

[3] http://project-modelino.com/english-phonetic-transcription-converter.php

3.4 The Automatic Media Transcription Module

This module is in charge of automatically inferring ypheme representations of speeches contained in audio or video content. It's implementation is possible using machine learning approaches that will use as training data the initial database of yphemes transcriptions that human contributors have recorded.

3.5 The Automatic Translation Module

Finally to implement the final opportunity of the Ynut system, we leverage the corpus built by contributors and enriched automatically by Ynut to provide automatic translation capabilities between spoken languages and written languages.

4 Concluding Remarks

We opened a discussion regarding a learning system for spoken languages. The main idea of the system is the enrichment of a phonetic database based on the phonetic alphabet of written languages. The linguistic contents may come from several sources such as audio/video databases and social networks. Then we have proposed an architecture for our system and we have presented its main functionalities and the research challenges for each functionality.

Further development will consider the implementation of a generic platform that may be used by any community that intend to promote its spoken language. Then we will consider the enrichment of phonetic data-base for the authors native spoken languages.

References

1. P. Austin and J. Sallabank. *The Cambridge handbook of endangered languages /*. Cambridge University Press„ Cambridge ;, 2011.
2. S. Besling. A statistical approach to multilingual phonetic transcription. *Philips Journal of Research*, 49(4):367 – 379, 1995.
3. D. Crystal. *Linguistics*. 1971.
4. P.-T. Nguyen, X.-L. Vu, T.-M.-H. Nguyen, V.-H. Nguyen, and H.-P. Le. Building a large syntactically-annotated corpus of vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 182–185, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
5. M. Ordean, A. Saupe, M. Ordean, M. Duma, and G. Silaghi. Enhanced rule-based phonetic transcription for the romanian language. In *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2009 11th International Symposium on*, pages 401–406, 2009.
6. J. Ouoba and T. F. Bissyandé. Leveraging the Cultural Model for Opportunistic Networking in sub-Saharan Africa. In *4th International IEEE EAI Conference on e-Infrastructure and e-Services for Developing Countries*, AFRICOMM, pages 1–10, Yaoundé, Cameroun, 2012.